DOCUMENT RESUME

ED 227 151 TM 830 162

AUTHOR Schmidt, William H.; And Others
TITLE Validity as a Variable: Can the Same Certification Test Be Valid for All Students?
INSTITUTION Michigan State Univ., East Lansing. Inst. for Research on Teaching.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO IRT-OP-53
PUB DATE Jul 82
CONTRACT 400-81-0014
NOTE 34p.
AVAILABLE FROM Institute for Research on Teaching, College of Education, Michigan State University, 252 Erickson Hall, East Lansing, MI 48824 ($3.25).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Competency Based Education; Elementary School Mathematics; Elementary Secondary Education; Evaluation Criteria; Graduation Requirements; Measurement Techniques; *Minimum Competency Testing; Models; *Student Certification; Student Evaluation; *Test Validity
IDENTIFIERS Florida Functional Literacy Test

ABSTRACT

Content, instructional, and curricular validity, as related to certification tests, are examined. All three deal with content validity but the domain differs among the the three. Certification tests must have instructional validity, i.e., the test must be valid both with respect to the domain used to define the minimum competencies and the instructional content domain (what is taught in the schools). The test items must be representative of the objectives domain but not necessarily representative of the instructional content domain. Whether a test has content validity with respect to the domain specified by the curricular materials is important only insofar as it is a surrogate for instructional validity. Curricular validity should not be used as a criterion to establish the instructional validity of a certification test. For tests of certification, a relatively large percentage of the items should represent topics that are covered by all students in the district and/or state (to assure that certification tests have instructional validity). A prototypic measurement of curricular and instructional validity for elementary school mathematics illustrates these points. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227151

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. C. Porter

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Institute for Research on Teaching

College of Education — Michigan State University

Occasional Paper No. 53

VALIDITY AS A VARIABLE: CAN THE SAME CERTIFICATION TEST BE VALID FOR ALL STUDENTS?

William H. Schmidt, Andrew C. Porter,
John R. Schwille, Robert E. Floden,
Donald J. Freeman

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

July 1982

This work is sponsored in part by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Program for Teaching and Instruction of the National Institute of Education, United States Department of Education. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-81-0014)

Institute for Research on Teaching

The **Institute for Research on Teaching** was founded at Michigan State University in 1976 by the National Institute of Education. Following a nationwide competition in 1981, the NIE awarded a second contract to the IRT, extending work through 1984. Funding is also received from other agencies and foundations for individual research projects.

The IRT conducts major research projects aimed at improving classroom teaching, including studies of classroom management strategies, student socialization, the diagnosis and remediation of reading difficulties, and teacher education. IRT researchers are also examining the teaching of specific school subjects such as reading, writing, general mathematics, and science, and are seeking to understand how factors outside the classroom affect teacher decision making.

Researchers from such diverse disciplines as educational psychology, anthropology, sociology, and philosophy cooperate in conducting IRT research. They join forces with public school teachers, who work at the IRT as half-time collaborators in research, helping to design and plan studies, collect data, analyze and interpret results, and disseminate findings.

The IRT publishes research reports, occasional papers, conference proceedings, a newsletter for practitioners, and lists and catalogs of IRT publications. For more information, to receive a list or catalog, and/or to be placed on the IRT mailing list to receive the newsletter, please write to the IRT Editor, Institute for Research on Teaching, 252 Erickson Hall, Michigan State University, East Lansing, Michigan 48824-1034.

Co-Directors: Jere E. Brophy and Andrew C. Porter

Associate Directors: Judith E. Lanier and Richard S. Prawat

Editorial Staff
Editor: Janet Eaton
Assistant Editor: Patricia Nischan

Abstract

The authors examine three types of content validity relative to certification tests such as the high school graduation test used in Florida. They identify and give examples of the domains for curricular, instructional, and content validity and make a case for the necessity of establishing the overall content validity of a test based upon the objectives that underpin the test before questioning whether that test is also valid with respect to the curricular material used in schools or even more narrowly with respect to the actual instruction provided in the schools. The authors describe and illustrate a prototypic measurement of curricular and instructional validity for elementary school mathematics.

VALIDITY AS A VARIABLE: CAN THE SAME CERTIFICATION TEST BE VALID FOR ALL STUDENTS[1]

William H. Schmidt, Andrew C. Porter, John R. Schwille, Robert E. Floden, and Donald J. Freeman[2]

In the judicial case of Debora P. vs. Turlington, the courts addressed the concept of validity as it pertained to the Florida Functional Literacy examination. Since the test was to be used in certifying a level of functional literacy required for high school graduation, much was at stake. Out of the controversy surrounding the examination and its use, two new types of validity emerged, curricular validity and instructional validity. The purpose of this paper is to explore the meaning of these two new types of validity, to show where they fit within the psychometrics tradition, and to consider what determines the curricular and/or instructional validity of a test.

Three Types of Content Validity

This conference is concerned not with validity in general, but more narrowly with the concept of content validity. The American Psychological Association (1974) defines content validity as the

---

[1]This paper was sponsored by the Ford Foundation and presented in October, 1981 at Boston College, Chestnut Hill, Massachusetts.

[2]William H. Schmidt is the director of the Language Arts Project and a member of the Content Determinants Project in the IRT. He is a professor and chairperson of educational psychology. Andrew C. Porter is director of the Content Determinants Project and co-director of the IRT. He is a professor of educational psychology and associate dean of program development. John Schwille, Robert Floden, and Donald Freeman are senior researchers at the IRT and members of the Content Determinants Project. All three are associate professors of teacher education.

situation in which the behaviors measured in a test constitute a representative sample of the behaviors to be exhibited in the desired performance domain. However, the case of Debora P. vs. Turlington raises a complication not addressed in this definition. For, while the lower court found that the test had reasonable content validity with respect to the skill objectives developed by the state board of education, the appellate court maintained that there was an additional question of curricular validity. The planners of this conference have added to the complexity by introducing still another term: instructional validity. What, we are asked, do these terms mean and what are their implications for tests of certification?

Defining the Three Terms

Most large scale testing programs such as state assessment, minimum competency, and certification tests have used a set of instructional objectives as "the desired performance domain" against which to judge content validity.

As for curricular validity, the judges seemed to be concerned with using the schools' curricular materials as the domain against which to judge a test. By the same token instructional validity can be defined as content validity with the domain of interest being the instructional content actually delivered by teachers in school.

The term content validity, as used by the trial court in Debora P. vs. Turlington, referred to the extent to which the test accurately reflected the domain specified for development of the test, namely the set of skill objectives defined by state legislation and acted upon by the state board of education. Curricular validity asks whether the test, established as valid with respect to the domain of

objectives, is also consistent with the curricular materials used in the school system where it is to be administered. Similarly, instructional validity is a matter of whether the test, however valid with respect to the objectives, adequately samples the instructional content actually taught to the students. In the discussion which follows, we refer to these three domains as the *objectives domain*, the *curricular materials domain*, and the *instructional content domain*.

Relationship of the Three Domains

If one were to think of each domain as a set, the interrelationships among the types of validity can be seen through Venn diagrams such as portrayed in Figure 1. If content validity were equated with validity for the objectives domain, as in the Florida case, the test must adequately sample subsets A, B, E, and F. However, subset A represents content in which students taking the test were not instructed, neither was that content included in the materials used by the schools.

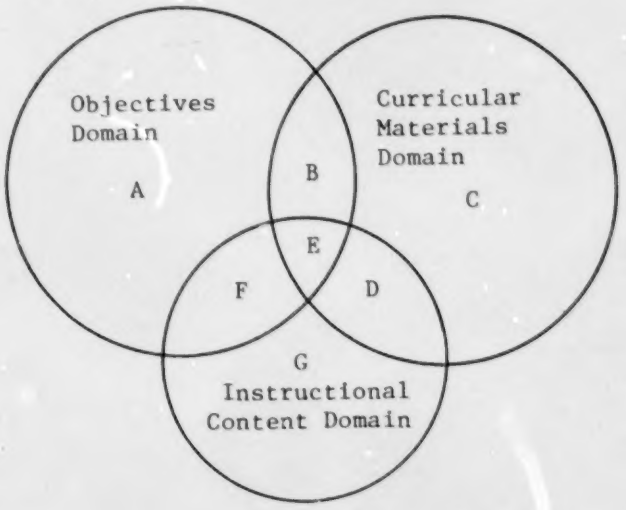


Figure 1. Interrelationships among the types of validity.

If a test has content validity with respect to both objectives and instructional content, then it is likely that the relationship shown in Figure 2 would obtain. In this case the objectives domain

is a proper subset of the instructional content domain. This arrangement seems reasonable inasmuch as certification tests are commonly based on minimal competencies. The scope of the instructional domain is large, reflecting its lack of restriction to minimal competencies. However, the domain of curricular materials need not be coincidental with either of the other two domains. Further, since different children may receive different content, it is quite conceivable that the model of Figure 2 would change across children within the same classroom as well as children in different classrooms.

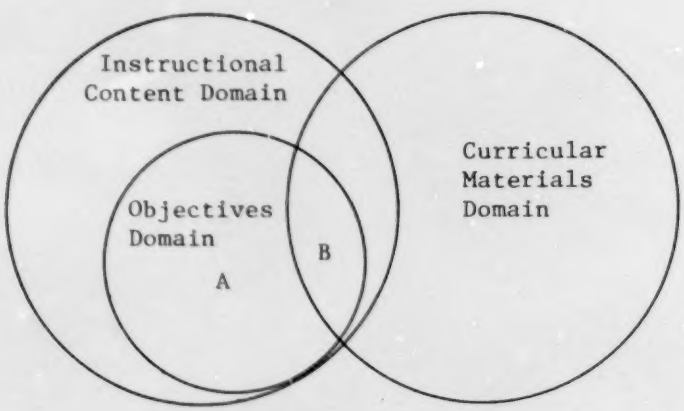


Figure 2. Interrelationships among types of validity.

If the test has content validity with respect to both objectives and curricular materials, then the objectives domain is likely a proper subset of the curricular materials domain, suggesting a situation similar to the one above in which a test of minimal competencies is being used.

Problems in Defining Domains

One of the problems with defining a domain concerns the level of detail to be contained in that domain. The domain should be at a level fine enough to make distinctions that are important but not to a level of detail so fine as to classify everything within the

subject matter as being different. This, of course, is the trick of being knowledgeable about 1) the subject matter and 2) the amount of transfer in learning that can occur among the topics contained in the domain. For if transfer of learning is straightforward between two topics (e.g., instruction on how to add 5 + 3 enables one to correctly do the problem 4 + 3), then a taxonomy that makes such distinctions might be overly detailed. On the other hand, it is obvious that at a very high level of generality, most all topics are similar (e.g., all items on a mathematics test deal with mathematics), so that moving in any direction too far is similarly not of any great value.

Another problem in specifying domains is tied more closely to the instructional content domain and the curricular materials domain. This is the question of topic emphasis: Is it sufficient for a topic to be included in the domain if it is covered in the school one time on one day or if it is found in one problem in the textbook. If this is not the case then what number of hours, days or problems is sufficient in order for the topic to be included in the domain?

Making the Test Representative

Figures 1 and 2 do not address one aspect of the traditional definition of content validity, namely, that the test be a *representative* sample of behaviors from the domain. When one considers the objectives domain alone, this property seems clearly desirable. Otherwise, one objective (e.g., a computation objective in mathematics) might be overemphasized to the detriment of another (e.g., an applications objective in mathematics). However, it is not so clear that the original motivation for introducing the terms curric-

ular validity and instructional validity are best served by retaining the requirement of representativeness. If we consider a test of minimal competencies, for example, the requirement for representativeness could be interpreted to mean that the content of the curriculum materials and the content of instruction must be limited to minimal competencies for the test to have curricular validity or instructional validity. Otherwise, there would be content in the materials and content in the instruction not represented on the test. To be restrictive in this way seems undesirable. The concepts of curricular validity and instructional validity serve in the eyes of the court and the planners of this conference to provide assurance that test content is also covered in curriculum materials and in class. The requirement for representativeness could change the concept of curricular and instructional validity from an assurance of sufficient coverage to a limitation on coverage.

If the requirement for representativeness is dropped, then in a strict sense, curricular validity and instructional validity cannot be thought of as specific types of content validity on a par with objectives validity. Rather, they should be thought of as characteristics of interest for tests that have first been judged to be valid with respect to the objectives domain (which could once again be equated with content validity). Since validity is a matter of degree rather than a dichotomous state, curricular validity and instructional validity would, in practice, need to be judged directly against the test rather than against the objectives domain.

On the other hand, the merit of requiring representativeness as a criterion for curricular or instructional validity is that this criterion would guard against a test giving too much weight to topics

that are minor or trivial aspects of instruction. For example, a test of minimal competencies might be devoted entirely to basic number facts. Would this test be considered to have curricular or instructional validity solely on the basis that they were covered in the materials or in the classroom even if other important aspects of the materials or classroom coverage were entirely neglected?

Thus, the question of representativeness seems to revolve around the issue of whether curriculum materials and classroom instruction are considered worthy indicators of content priorities in their own right or, alternatively whether the objectives domain is considered a sufficient criterion of content priorities, with the curriculum materials and classroom instruction being taken, not as indicators of content priorities, but of sufficient student opportunity to learn. There may be no general answer to this question since in part it is dependent upon the extent to which the objectives are viewed as authoritative. Presumably the greater the overlap among the three domains, the more authoritative each would be viewed as a guide for what should be tested.

A discussion of our attempts to measure the overlap between tests, curriculum materials, and classroom instruction, which follows later in this paper, will serve to further illustrate these issues. Given the general nature of state assessment, minimum competency and certification tests, and the issues before the courts, the answer for these types of tests seems to be the latter (i.e., to not require representativeness for a test to have curricular and instructional validity).

What Type of Validity for What Type of Test?

For general aptitude or general achievement tests we would argue that the main concern should be content validity with respect to the domain upon which the test is to be based.

For tests of certification, it is not enough that a test have validity with respect to the objectives domain. It should also have content validity with respect to instructional content but not necessarily in a representative fashion. In other words, some acceptably large percentage of the items sampled from the objectives domain must also be covered by every student in every classroom (Figure 2). This is the issue of sufficient student opportunity.

If tests without instructional validity are being used for certification, the students who fail such tests are being penalized for the failures of the schools and teachers and not for their own inadequacies. The rational basis for judging student performance in school is undermined.

If a test has instructional validity, curricular validity can be argued to have little importance, to be superfluous. In fact, at least two arguments can be made for curricular validity: One is that curriculum materials can serve to reinforce classroom coverage of all the content on the test; the other (to be developed in the next sections) is that it is more difficult to measure instructional validity than it is to measure curricular validity. There is the possibility of using curricular validity as a surrogate for instructional validity, and it is relatively easier to control curricular validity than instructional validity.

Prototype Measurement of Curricular and Instructional Validity

In this section we set forth a system to be considered for use in content validation. Suggestions here are the result of work in elementary school mathematics. It is our hope that some of what we have learned in this context might be generalizable to other subject matters and to other grade levels as well.

A Taxonomy for Measuring Content Validity

In our research on the determinants of content coverage in the classroom, we were interested in developing an instrument that would enable us to measure the content of instruction, tests, and curricular materials. It is our proposal that such a device could also be used to establish the content validity of a test with respect to any of the domains discussed in this paper. A taxonomy that enables one to map the items of a test into their content specifications for fourth-grade mathematics could be used to characterize the content domains represented by that test. This taxonomy could also be applied to the other domains. For example, the domain specified by the objectives on which the test is to be constructed could be analyzed by content using this taxonomy. Since this could also be done for the tests, a way to establish the content validity of the tests is to determine the degree to which the test item map can be subsumed under the objective map.

This same strategy could be followed with respect to curricular materials. The various curricular materials could similarly be analyzed by content using the taxonomy, and a map could be devleoped that suggests the range of topics represented in the domain covered by the textbook or other curricular materials. The same thing

could be done with respect to the content of the actual classroom instruction. In a later section we address the additional problem of how one takes the actual classroom instruction and maps that into the taxonomy.

Description of a Taxonomy for Elementary School Mathematics

The taxonomy discussed here takes the form of a three-dimensional matrix. The three dimensions are: the general intent of the lesson (e.g., conceptual understanding or application), the nature of the material presented to students (e.g., measurement or decimals), and the operation students must perform (e.g., estimate or multiply). Developed in conjunction with this taxonomy is a set of rules to operationalize the cell boundaries. The application of the taxonomy to tests and textbook exercises is relatively straightforward, susceptible to being replicated, and results in high inter-rater reliability (Freeman et al., 1981).

Application of the Taxonomy to Tests

Each item on the test is examined and classified according to the taxonomy. The data from such an analysis can be represented by a mark on the taxonomy that indicates which of the cells in the taxonomy are covered. After the entire test has been mapped onto the taxonomy, the result is a visual representation of the areas covered by that test. This process is illustrated in Figure 3, which portrays the results of the content analysis of the Stanford Achievement Test (SAT). It illustrates the flexibility of the taxonomy to describe content at different levels of detail. Specific topics are represented by the cells of the classification matrix (e.g., three of the 112 Stanford items focus on the skill of column addition of multiple

SAT - Intermediate Level I

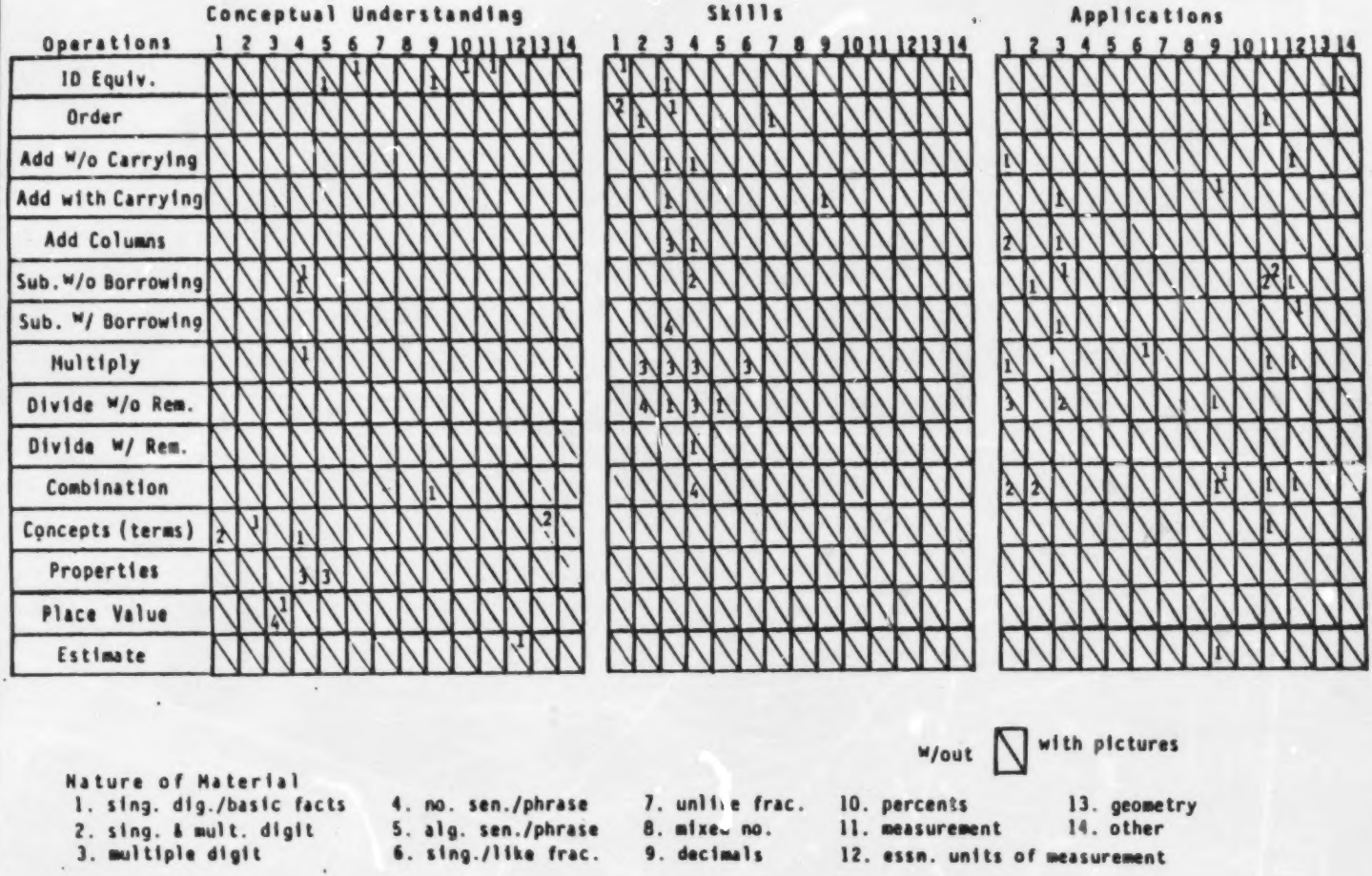


Figure 3. Content analysis of Stanford Achievement Test (Intermediate Level/Grades 4.5-5.6), 1973

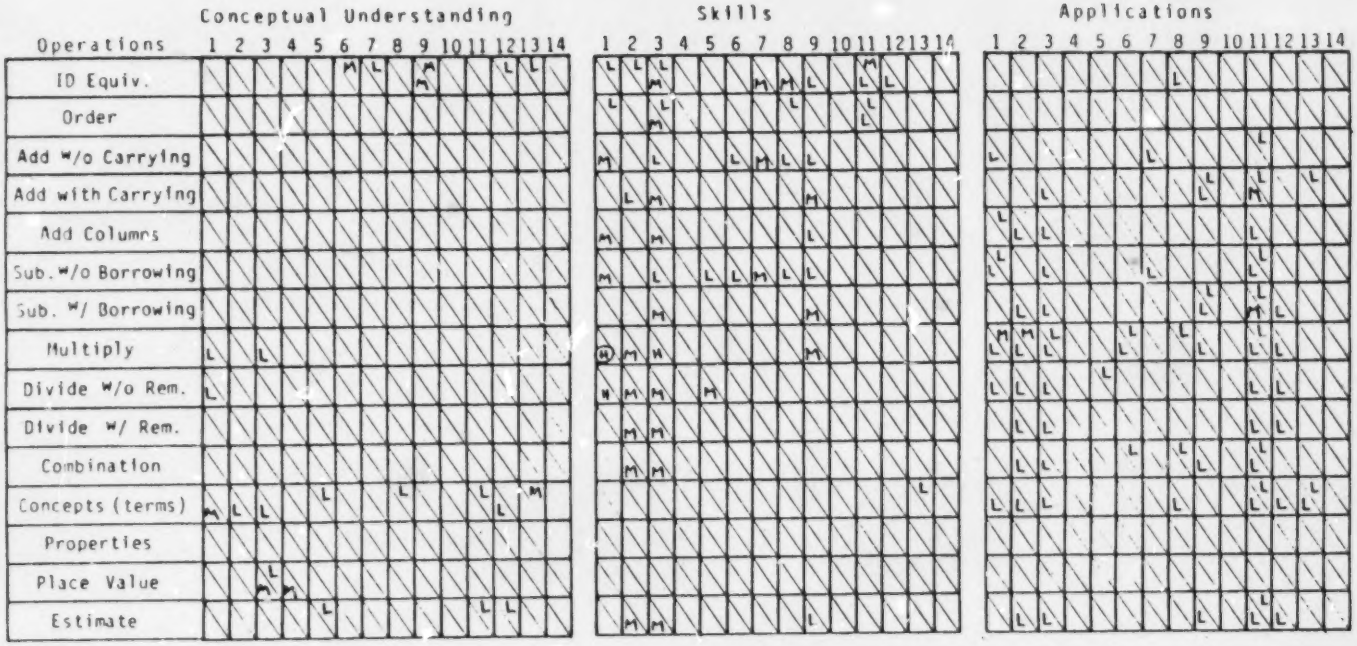
digit numbers). More general topics can also be addressed by summing across cells to obtain marginal totals (e.g., seven of the 112 items deal with column addition).

Application of the Taxonomy to Curricular Materials

The use of the taxonomy to analyze the content of curricular materials is much more difficult than it is for tests. Lessons in textbooks contain two distinct components: instructional activities directed by the teacher and practice exercises assigned to students. Our analyses of textbooks were limited to items in the student exercise portion of each lesson. The number of items to be classified for the student exercise portions of the three textbooks that we have worked with range from a low of 4,288 items in the Addison-Wesley textbook to a high of 6,968 items in the Houghton-Mifflin text. These figures show that the content classification of curricular materials such as textbooks is extensive and time consuming.

To illustrate the application of the taxonomy to curricular materials, we provide the results from the content analysis of three fourth-grade textbooks: Mathematics in Our World, Addison-Wesley Publishing Co., 1978; Mathematics, Houghton-Mifflin Co., 1978; and Mathematics Around Us, Scott-Foresman and Co., 1978.

An analysis of content at the cell level within the taxonomy provides a basis for comparing the treatment of specific topics within textbooks (e.g., applications involving the multiplication of single-digit numbers). Figure 4 depicts the concentration of items representing specific topics within one of the three texts. Four general categories are used to depict the relative frequency of items in each



L = less than 0.5% (<21 Items)
M = 0.5% to 5.0% (21-214 Items)
H = 5% to 10% (215-429 Items)
(H) = More than 10% (> 429 Items)

w/out ▨ with pictures

Nature of Material

1. sing. dig./basic facts
2. sing. & mult. digit
3. multiple digit
4. no. sen./phrase
5. alg. sen./phrase
6. sing./like frac.
7. unlike frac.
8. mixed no.
9. decimals
10. percents
11. measurement
12. essn. units of measurement
13. geometry
14. other

Figure 4. Distribution of items in the Addison-Wesley fourth-grade text

cell of the taxonomy. The symbol "H" denotes high frequency cells into which 5% to 10% of the items fall. When the "H" is circled, over 10% of the items in the text are concentrated in that cell. An "M" designates cells containing a moderate concentration of items (0.5% to 5.0%) and "L" indicates cells containing a low frequency of items (less than 0.5%). Cells that have no symbols are "empty," meaning that content does not occur in the textbook.

Each textbook's distribution of specific topics across the categories of concepts, skills, and applications is presented in Table 1. From this table it can be seen that of the 293 topics included in one or more of the three books, 51% were included in the Addison-Wesley text, 57% in the Houghton-Mifflin text, and 67% in the Scott-Foresman text. Because there was overlap in topic coverage (i.e., some topics were covered in two or three of the books), the cell frequencies in Table 1 sum to more than 293. Nevertheless, this analysis reveals that any given book covers only a little more than half the topics presented in all three books collectively.

Table 1.

Distribution of Specific Topics Across Concepts, Skills and Applications

	Addison Wesley	Houghton Mifflin	Scott Foresman
Concepts	23	42	56
Skills	53	52	66
Applications	72	73	75
Totals	148	167	197

The Establishment of Curricular Validity: Textbooks

In this section we illustrate the way in which curricular validity can be established by examining the map between the several tests and textbooks illustrated in the previous section. In order to put this in the context that we have been considering, assume that the test's content validity with respect to the objectives domain has already been established and that what is being considered here is the additional question of the degree to which the test has curricular validity with respect to the textbook being used in that particular district.

For purposes of illustration, data are presented that contrast the content domains specified by (1) the five most frequently used standardized tests in mathematics and (2) three textbooks. With our data we can ask what percent of the topics on a test are covered in a given textbook. The four columns labeled "T" in Table 2 describe the percent of topics in each test that served as the focus of at least one item in the student exercises in each book. In interpreting these figures, it is important to remember that at least 4,000 items were classified for each book. The percent of tested topics covered in a given book ranged from a low of 52.8% for the SAT and Houghton-Mifflin text to a high of 73.7% for the MAT and Houghton-Mifflin textbook. Thus, only about one-half of the topics that were considered in the SAT SAT were covered by one or more of the 6,986 items in the student exercise portions of the Houghton-Mifflin text.

The columns labeled "T" in Table 2 describe the percent of test topics that served as the focus of at least 20 items in each book. If one assumes that this subset of book topics represents the content students will have had an adequate opportunity to learn or to prac-

Table 2.

Percent of Tested Topics Covered in Each Textbook

Tests	Publisher					
	Addison-Wesley		Houghton-Mifflin		Scott-Foresman	
	T[a]	T'[b]	T	T'	T	T'
	(148)[d]	(42)	(167)	(49)	(197)	(50)
MAT (38)[c]	63.2	31.6	73.7	39.5	73.7	42.1
SAT (72)	54.1	22.2	52.8	20.8	62.5	22.2
IOWA (66)	54.5	25.8	72.7	31.8	71.2	25.8
CTBS I (53)	56.6	32.1	64.2	37.7	64.2	35.8
CTBS II (61)	60.7	27.9	59.0	37.7	67.2	34.4

[a] T = Topics covered by at least one item in the book.

[b] T' = Topics covered by at least 20 items in the book.

[c] Numbers in parentheses indicate the total number of topics in each test that are covered in all three books.

[d] Numbers in parentheses across textbooks indicates the number of items in each textbook that are covered in all tests.

tice during the preceding academic year, these figures should provide reasonable estimates of the relation between test content and the content of instruction suggested by the book. These values ranged from a low of 20.8% for the SAT and Houghton-Mifflin text to a high of 42.1% for the MAT and Scott-Foresman text. In other words, the proportion of topics presented on a standardized test that received more than cursory treatment in each textbook was never higher than 50%!

The Establishment of Curricular Validity: District Objectives

Still another example of an examination of curricular validity is presented with respect to the mathematical objectives used in a district, which we call Knoxport. The full strand of mathematical objectives, excluding those dealing with enrichment, was subjected to a content analysis. This mapping of the objectives was then contrasted with the Stanford Achievement Test, which also happens to be the standardized test administered in that district.

The content specified by the objectives is not totally covered on the Stanford Achievement Test, nor, for that matter, are the topics tested on the Stanford Achievement Test all present in the district objectives. There is a fair amount of overlap between the two sources but this is in no sense complete.

One way to suggest this comparison is in terms of the objectives used by the school district. Of the total number of objectives, 56 percent have content that is tested on the Stanford Achievement Test. These 52 objectives, however, do not represent distinct topics as defined by the taxonomy. In fact, the 52 objectives are classified into 24 cells of the taxonomy. Another way to think of this lack of consistency is to point out that 44 percent of the topics covered

by the district objectives are topics that are also tested by the SAT. Either figure implies that only one-half of the content covered by the objectives is tested by the Stanford Achievement Test.

Another way to look at the lack of consistency between the two is to consider it from the other point of view. The items on the SAT represent 61 cells or topics in terms of the taxonomy. Again remembering that 24 topics are covered by both the district objectives and the SAT implies that approximately 40 percent of all topics covered by the SAT are also similarly covered in terms of the district objectives. From this perspective, there is even a greater discrepancy. The SAT items deal with many topics not covered by the district objectives.

Establishment of Instructional Validity

The application of the taxonomy to instructional content is a much more difficult task. Tests and curricular materials are almost always expressed in written form and hence are rather easily subjected to a content analysis using the taxonomy. The content of instruction, however, is more elusive as it represents an on-going process that is presented to the students interactively with the teacher. Obtaining data on instruction and detailing the content of that instruction is a difficult task.

At the Institute for Research on Teaching (IRT) we have used various forms for the collection of such information. The most costly is field observation. In this approach, trained observers record during the course of the day what topics are covered and for what periods of time. A cheaper and more straightforward approach to the problem is to have teachers keep daily logs in which they record

the content of their instruction.

An important question is the degree to which logs kept by teachers are an accurate representation of topic coverage. In one project at the IRT, we found that teachers in general were able to keep fairly accurate logs. An analysis of the measurement error inherent in the process is being conducted and preliminary results suggest that in the aggregate (that is averaged over days), the amount of error is using the logs to represent content/time allocations of teachers is not unacceptably large.

Hence, the instructional validity of a test can be established here as was illustrated with respect to curricular validity. The only difference being that within this context the content profile of the test is contrasted with the content profile derived from the log analysis. The degree to which the two are consistent with each other is the extent to which instructional validity is present for the test and for the students in that particular class. None of the data from our study were in a complete enough form to provide us with an illustration contrasting content coverage against one of the standardized tests. We can, however, discuss, in general, examples of teachers who used the same materials but whose content coverage varied.

During the 1979-80 school year we collected extensive data on seven teachers in three different districts. We interviewed the teachers weekly, observed their classroom instruction, and had them keep daily logs recording their fourth-grade mathematics instruction.

In the Sawyer district we observed two teachers whom we shall call Wilma and Jacqueline. The Sawyer district had a mandated

mathematics textbook series (Holt) that all teachers were required to use. However, teachers were not told they had to teach all topics from the book. In order to place the observations made in this study into the context of this paper, imagine that this district is considering a minimal competency test for promotion from fourth to fifth grade and that we are concerned with the mathematics section (a totally hypothetical situation). Let us further assume that the superintendent and curriculum director have specified the domain upon which this test is to be developed and that, through a careful analysis of the Holt text, have decided that the domain specified by the objectives is a proper subset of the domain of topics generated by the Holt textbook: In other words the test has curricular validity. But would it also have instructional validity? One might think that having established curricular validity and also having a standardized textbook, so as to assure curricular validity for all students in the district, would assure that all students would receive instruction on every topic contained in the domain specified by the objectives. In other words, this would insure instructional validity. However, in Sawyer the two teachers we observed treated the textbooks in very different ways. For Jacqueline, the textbook essentially defined--for this particular year at least--the content of her instruction. She followed the textbook in an almost linear fashion covering it page by page until she ran out of time at the end of the year (at Chapter Nine). A test such as the hypothetical promotion test suggested above would have had instructional validity for Jacqueline's classroom if it contained the same content as the first nine chapters of the textbook.

Two caveats need mentioning. Three students in Jacqueline's classroom were put into a special subgroup that used the third grade Holt book because these students were below grade level. Obviously a mathematics test that matched the fourth grade text would not have had instructional validity for these students. It is also interesting to note that if the material covered in the test was not concentrated at the beginning of the textbook but was found throughout the textbook then the issue of how far the students went in the textbook does determine whether the test would have instructional validity. In other words, if the test examines domain topics covered in the back sections of the book then Jacqueline's students would not have been instructed in them and the certification tests would not have been instructionally valid for those students.

The other teacher in this district, Wilma, did not follow the textbook in any straightforward fashion. In fact, this teacher had her own conception of what should be covered in fourth-grade mathematics. This conception not only included a detailing of the topics that should be covered but also a time schedule as to when these topics should be covered. As a result of this, Wilma did not cover the textbook. She rearranged the order in which she covered things in the textbook; skipped sections of the book that she did not find consistent with her own conception of what should be covered, and added to the instruction topics that were not contained in the book.

In this case, it is clear that although the textbook was mandated, the teacher chose to use it in her own fashion. If any of the topics that she chose to skip were a part of the domain on which the test was based, then, despite the fact that the test had curricular validity, it would not totally have had instructional validity for

the students in Wilma's class. In general, our data show that students in this district using the same mandated textbook in mathematics received different instructional content.

The implication of this is that the hypothetical promotion certification test would have had content validity with respect to both the objectives and the curricular material for all students in the district but for the students in Jacqueline's class the test would additionally have had better content validity with respect to the domain of instructional content than would have been the case for students in Wilma's class.

Consider one last example. In Knoxport a detailed strand of objectives for mathematics was required for use by all teachers. Associated with this set of mathematical objectives was a management system that included locator tests, pretests, and mastery tests. Teachers kept records on the objectives that students had passed. In fact, although to the best of our knowledge it was never invoked, a policy existed whereby teachers could be released from their jobs if they did not use the MBO system and have the students in their class work through the objectives. It is interesting that even in this district with paper sanctions for not following the system, we found, among the three teachers studied in this district, a lack of consistency in terms of their students covering the objectives. One of the teachers, Andy, almost totally followed the MBO system and had his students work systematically through the objectives, one by one, until they passed the mastery tests. For students in his class any test for advancement made consistent with the objectives would have been valid with respect to instructional content at least for some students but would have varied student by

student since not all were able to progress through all the objectives for that grade level due to self pacing.

However, in the same district two other teachers, Terry and Lucy, followed less closely the MBO system for their mathematics instruction. Lucy provided two mathematics sessions, one devoted to regular mathematics instruction and the other, devoted to using the individualized objectives system. The other teacher, Terry, rarely used the MBO system, and, in fact, by the end of the year, the students had spent very little time in the system. Although students in Terry's class were from the same district, their testing would not necessarily have been valid in content even if it were consistent with district objectives.

The Three Types of Content Validity and Implications for Curriculum Policy Making

Studies we have done indicate that if no efforts are made to assure curricular or instructional validity, a test that has content validity with respect to the objectives domain would vary in its curricular and instructional validity for different students. Consider for example a test that has a curricular validity (i.e., the test has content validity both with respect to the domain of objectives and with respect to the domain defined by the curricular materials). A test will not have this characteristic unless the materials have been standardized for the population being tested, (e.g., the state or district). Otherwise one must talk about curricular validity in relationship to some district, school, or building. In this way, validity becomes a variable and is not a constant characteristic of the test itself as it is in classical test theory and in the case of content validity based on a set of objectives. In general, for

curriculum to be valid, the curricular materials must be uniform among the population for which the test is designed. For example, statewide adoptions of textbooks would assure that a test based on the objectives and consistent with the textbook would have content validity both with respect to the objectives domain and the curricular materials domain.

Many educators assume that all basic textbooks in a certain subject matter area cover the same basic content and are in fact interchangeable. This would imply that the test would have curricular validity with respect to any one of these textbooks. In the work we have done with three fourth-grade mathematics textbooks including Scott-Foresman, Houghton-Mifflin, and Addison-Wesley, we found substantial differences among them, which implies that these books are not interchangeable with respect to their definition of a curricular domain. One cannot assume that any book within a certain subject matter will guarantee curricular validity. Once the content domain with respect to the objectives is specified, careful analysis of the major textbooks in the field must be undertaken so as to guarantee that the content domain specified by the objectives is in fact coincidental or at least a subset of the domain defined by the curricular material.

At this point it is reasonable to ask the question, why anyone would be particularly concerned about a test having curricular validity? One reason is the belief by many educators that the materials do in fact specify the actual instruction to take place in the classroom (i.e., by assuring that a test has curricular validity you are also simultaneously assuring instructional validity).

Also on the practical side, policies that insure curricular validity are more easily established than is the case for instructional

validity. For example, it is relatively straightforward for a district superintendent or state superintendent to mandate textbooks or curricular materials to be used in the schools within that unit. This is not the case for mandating the actual content of instruction. It is also easier to establish whether a test has content validity with respect to the curricular materials domain. The establishment of instructional validity is much more difficult and time consuming.

But we have found that even when materials are required to be used by all teachers within a district or building, this does not guarantee that what is in those curricular materials will necessarily be covered in every classroom. Many teachers operate relatively autonomously in defining the content of their instruction. This we at least found to be the case for fourth-grade mathematics. Some teachers follow textbooks and other curricular materials almost to a tee, whereas other teachers in those same districts and under the same mandates will not necessarily cover nor follow the textbooks. Consider the case of Jacqueline and Wilma as reported previously. It appears to us that verifying the consistency between test items and the curricular domain does not insure consistency between test items and the instructional content domain. Since the latter is the desired standard, for former would only be useful when it could serve as a surrogate for the latter. The research we have done certainly challenges the expectation that this would occur frequently.

How could curricular validity serve as a reasonable surrogate for instructional validity? If management systems such as the MBO system used in Knoxport were to have associated with them stringent rewards and punishments that assured that all children will cover

the objectives, and if a test has curricular validity with respect to the objectives, this system might guarantee instructional validity.

Why Does Instructional Validity Vary?

One might ask why all students in the same classroom do not receive identical instructional content. Two reasons are suggested by our research. The first pertains to grouping strategies. If the class is always taught as a whole group then all children within that classroom will receive the same basic content. Hence for this situation a test has instructional validity for all students within the classroom.

If the instruction within the classroom is provided on a subgroup or individual basis, identical instructional content is not necessarily assured across all individuals or subgroups of students. This would imply that, even within the same classroom, a test might have instructional validity with respect to one subgroup but not with respect to others.

Many of the certification or diploma tests measure cumulative types of educational experiences. A second reason why instructional validity is not guaranteed for all students in the same classroom is that content, assumed to have been covered in a previous grade level and hence not covered in the present grade, might not have been covered for all individuals (e.g., because of the classroom from which they came). So, for some students, certain content is not covered. To the extent that this happens, it exacerbates the problem of guaranteeing instructional validity for all children.

The point of this section is that instructional validity will not occur naturally. One way to encourage instructional validity is to require that certain curricular materials, consistent with the

domain specified by the objectives, be used and that sanctions be included so that teachers are more likely to cover those objectives using the instructional materials provided. One also wonders if the long-held notion of teaching to the test might have a positive effect in encouraging instructional validity.

When the test if first administered (and assuming it is valid with respect to the objectives), one cannot necessarily expect the objectives domain to be a subset of the instructional content domain for all students unless one puts some constraints on what is taught. A reasonable constraint is to require some level of performance on the test as a criterion for graduation. Requiring that this test have instructional validity before it can be used (as some have argued) is like a "Catch 22" since instructional validity is only likely to occur after such a testing practice has been in place for a while.

If a test used for certification is administered for several years prior to the time it will actually be used for certification, and if decisions and careful content analyses of the objectives domain (on which the test is based at the level of detail suggested by our taxonomy) is made available to the teachers, it seems likely that teachers would begin to teach to the test, which would provide for greater instructional validity.

Another way of insuring instructional validity is to give the test initially as a diagnostic device and then give remediation to students on the topics they fail. This in fact is pretty much the way the New York Regents Competency Test is supposed to work. The test is first given in 9th grade. Students who fail are put in a

special help class. They can take the test as many times as needed to pass.

Summary

In this paper we have examined the three concepts: content validity, instructional validity, and curricular validity. All three deal with content validity but the domain differs among the three. We maintain that certification tests must have instructional validity, by which we mean that the test must be valid both with respect to the domain used to define the minimum competencies and the instructional content domain (i.e., what is taught in the schools). We further argue that the test items must be representative of the objectives domain but not necessarily representative of the instructional content domain.

Whether a test has content validity with respect to the domain specified by the curricular materials is important only insofar as it is a surrogate for instructional validity. Some might believe that an analysis of the curricular materials tells us what content is covered in the schools. Our work suggests that this is far from true. Teachers in the United States generally operate fairly autonomously as decision makers in defining the content of their instruction. They are influenced by many sources other than curricular materials such as tests, principals, and other teachers. It is for this reason that curricular validity should not be used as a criterion to establish the instructional validity of a certification test. For tests of certification there must be some other way to assure that a relatively large percentage of the items represent topics that are covered by all students in the district and/or state (i.e., to assure that certification tests have instructional validity).

References

American Psychological Association. Standards for educational and psychological tests. American Psychological Association, 1974.

Freeman, D., & Belli, G. Influence of differences in textbook use on the match in content covered by textbooks and standardized tests. Paper presented at the National Council on Measurement and Education, Annual Conference, Los Angeles, April, 1981.